Bayesian treatment

Recap :-

* Our parameter estimation procedure involved :

$$\hat{\theta}_{ML} \leftarrow \underset{\theta}{\operatorname{argmax}} \log P(X|\theta)$$

* We used direct and indirect (EM) optimization

* We obtained point estimates of 'θ'

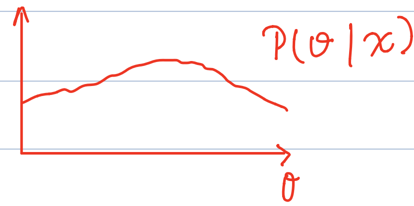
* What if we are interested in $P(\theta|X)$ instead?

Why $P(\theta|X)$?

* Prediction using full information

$$P(z_{new} \mid x_{new}, \mathcal{X}_{Train}) = \int P(z_{new} \mid x_{new}, \theta, \mathcal{X}_{Train}) \cdot P(\theta \mid x_{new}, \mathcal{X}_{Train})\, d\theta$$

e.g. cluster-membership indicator for a new data point



* With a point estimate, the best we can do is

$$P(z_{new} \mid x_{new}, \theta_{POINT}, \mathcal{X}_{TRAIN})$$

* "Integrating over" θ using $P(\theta \mid \mathcal{X})$ utilizes all the uncertainty/information about θ

How to compute $P(\theta \mid \mathcal{X})$?

Bayes' rule :- $P(\theta \mid \mathcal{X}) = \dfrac{P(\mathcal{X} \mid \theta)\, P(\theta)}{P(\mathcal{X})}$

Looks simple!
But can be really hard for many practical problems.

$$P(\theta|x) = \frac{\overbrace{P(x|\theta)}^{\text{Likelihood}}\overbrace{P(\theta)}^{\text{Prior}}}{\underbrace{\int P(x|\theta)P(\theta)\,d\theta}_{P(x)}}$$

Biggest Problem

* Normalization constant :- $\int P(x|\theta)P(\theta)\,d\theta$

* A subset of 'Likelihood, prior' pair, allows computing the normalization constant analytically.

* For example :- Multinomial distribution

$$Pr(m|\theta, N) = \binom{N}{m_1 \dots m_k} \prod_{k=1}^{K} \theta_k^{m_k}.$$

Where :-

$$m := [m_1, \dots, m_k]$$

$$\sum \theta_k = 1, \quad \theta_k > 0 \quad \forall k$$

$$Pr(\theta|m, N) = \frac{Pr(m|\theta, N) . P(\theta)}{\int P(m|\theta, N) . P(\theta)\,d\theta}$$

$$= \frac{\left[\binom{N}{m_1 \ldots m_k} \prod_{k=1}^{K} \theta_k^{m_k}\right] P(\theta_1, \ldots, \theta_k)}{\displaystyle\int_{\substack{\theta_1 \ldots, \theta_k > 0 \\ \sum \theta_k = 1}} \binom{N}{m_1 \ldots m_k} \prod_{k=1}^{K} \theta_k^{m_k} P(\theta_1, \ldots, \theta_k) d\theta_1 \ldots d\theta_k}$$

* I know of one $P(\theta)$ which would let us compute normalization constant (with an added benefit (we'll see)

$$P(\theta_1, \ldots, \theta_k \mid \underbrace{\alpha_1, \ldots, \alpha_k}_{\text{hyper-parameters}}) = \underbrace{\frac{\overbrace{\Gamma}^{\text{gamma function}}\left(\sum_k \alpha_k\right)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}}_{\text{Very similar to likelihood}}$$

gamma function ← continuous analog of factorial

So denominator,

$$\binom{N}{m_1, \ldots, m_k} \frac{\Gamma\left(\sum_k \alpha_k\right)}{\prod_k \Gamma(\alpha_k)} \overbrace{\int \prod_k \theta_k^{m_k + \alpha_k - 1} d\theta_1 \ldots d\theta_k}^{\text{using Beta Integral result}}$$

$$= \binom{N}{m_1 \ldots m_K} \frac{\Gamma\left(\sum_k \alpha_k\right)}{\prod_k \Gamma(\alpha_k)} \cdot \frac{\prod_k \Gamma(\alpha_k + m_k)}{\Gamma\left(\sum_k (\alpha_k + m_k)\right)}$$

So

pseudo-counts

$$P(\theta \mid m, N, \alpha_1, \ldots, \alpha_K) = \underbrace{\frac{\Gamma\left(\sum_k (\alpha_k + m_k)\right)}{\prod_k \Gamma(\alpha_k + m_k)} \prod_k \theta_k^{m_k + \alpha_k - 1}}_{\text{Same form as the prior}}$$

* Hence, we can further restrict our search for prior distributions by asking one question:-

Prior x Likelihood = Posterior

Same functional form.

* Such priors are said to be "conjugate" to likelihood

* In above example, $P(\theta|\alpha)$ is a dirichlet distribution - It is conjugate to Multinomial likelihood.

* For a family of distributions, such 'easy' priors exist - We'll generalize our treatment to that entire family (exponential family)

<u>Comment</u> :- Strict adherence to conjugate priors make Bayesian treatment very easy

- However, it is a major criticism on Bayesian methods too!